Guided Capstone Project Report

Big Mountain Resort is searching for explanatory variables, preferably which they can influence, that maximizes revenue and profit from ski resort tickets. The guided capstone narrowed the problem to being able to predict the AdultWeekend variable, which is the cost of an adult weekend chairlift ticket.

Cleaning and Pre-processing

Initial outliers were visualized with boxplots and removed by using a Boolean filter for values outside of 1.5 times interquartile range.

A heatmap was used to investigate and remove highly collinear variables with Pearson correlation coefficient > 0.95. Base elevation was the only removal.

	Unnamed: 0	summit_elev	vertical_drop	base_elev	trams	fastEight	fastSixes	fastQuads	quad	triple
Unnamed: 0	1.000000	-0.200000	-0.080000	-0.210000	-0.070000	-0.010000	-0.050000	-0.110000	0.090000	-0.090000
summit_elev	-0.200000	1.000000	0.740000	0.980000	0.340000	0.100000	0.260000	0.450000	-0.010000	0.180000
vertical_drop	-0.080000	0.740000	1.000000	0.590000	0.590000	0.180000	0.380000	0.680000	0.140000	0.260000
base_elev	-0.210000	0.980000	0.590000	1.000000	0.230000	0.070000	0.210000	0.340000	-0.060000	0.150000
trams	-0.070000	0.340000	0.590000	0.230000	1.000000	0.080000	0.530000	0.660000	0.220000	0.270000
fastEight	-0.010000	0.100000	0.180000	0.070000	0.080000	1.000000	0.150000	0.100000	0.090000	0.190000
fastSixes	-0.050000	0.260000	0.380000	0.210000	0.530000	0.150000	1.000000	0.440000	0.120000	0.270000
fastQuads	-0.110000	0.450000	0.680000	0.340000	0.660000	0.100000	0.440000	1.000000	0.150000	0.300000
quad	0.090000	-0.010000	0.140000	-0.060000	0.220000	0.090000	0.120000	0.150000	1.000000	0.140000
triple	-0.090000	0.180000	0.260000	0.150000	0.270000	0.190000	0.270000	0.300000	0.140000	1.000000

Model 1

The results of the 1st model had an extreme mean absolute error. The poor results are attributable to mainly two predictions, which are orders of magnitude off from the y_test values.

```
array([ 5.12800391e+01,  4.86023658e+01,  5.62733252e+01,  6.39972785e+01,
        5.68672584e+01,  6.95870154e+01,  5.73798316e+01,  4.09970725e+01,
        5.03526710e+01,  6.49967139e+01,  4.96184943e+01,  6.33552039e+01,
        7.07734168e+01,  5.10032142e+01,  4.39109290e+01,  4.23200782e+01,
        4.12041724e+01,  3.38913672e+01,  7.50949195e+01, -2.36904140e+07,
        4.16659034e+01,  4.50463508e+01,  6.31275149e+09,  3.91864112e+01,
        6.64172156e+01,  5.85417273e+01,  5.80561622e+01,  5.19921058e+01,
        3.93822730e+01,  6.85872596e+01,  3.25440772e+01,  5.78434852e+01,
        6.60809119e+01,  7.46625007e+01,  5.96816771e+01,  3.89713233e+01,
        4.91956161e+01,  3.39361747e+01,  5.73449805e+01,  7.76082862e+01,
        5.90939429e+01,  4.17928641e+01])
```

The source of the error for these predictions seems to be the scaled dummy variables for state. After data cleaning, there was only one row for Maryland and one row for New Jersey. This caused these features to have much more extreme values when scaled because the average and variance was very close to zero, making a row with a “1” many standard deviations above average.

state_Maryland	state_Massachusetts	state_Michigan	state_Minnesota	state_Missouri	state_Montana	state_Nevada	state_New Hampshire	state_New Jersey	y_pred	y_test
-0.078086881	-0.157622081	-0.316227766	-0.225733059	-0.110769755	-0.136082763	-0.078086881	-0.267261242	12.80624847	(23,690,414)	64.16681
12.80624847	-0.157622081	-0.316227766	-0.225733059	-0.110769755	-0.136082763	-0.078086881	-0.267261242	-0.078086881	6,312,751,492	79

```
Explained Variance Score: -4875532960905248.0
Mean Absolute Error: 150867669.3468477
```

Model 2

Eliminating the dummy state variables solved the major issue. The top 10 most important coefficients and model results are shown.

	Coefficient
AdultWeekday	1.093827e+01
clusters	3.366601e+00
summit_elev	2.283985e+00
projectedDaysOpen	1.838975e+00
daysOpenLastYear	1.715290e+00
triple	1.613914e+00
averageSnowfall	1.437612e+00
quad	1.404888e+00
surface	1.336300e+00
vertical_drop	1.059412e+00

```
Explained Variance Score: 0.7631640296668699
Mean Absolute Error: 5.726953545349737
```

Model 3

As instructed by the guided capstone directions, summit elevation was removed because it cannot be adjusted by management. The main difference is that it ranks 'average Snowfall' as a more important coefficient.

	Coefficient
AdultWeekday	1.011003e+01
averageSnowfall	3.314421e+00
clusters	2.652311e+00
daysOpenLastYear	2.414098e+00
vertical_drop	1.692971e+00
projectedDaysOpen	1.466963e+00
triple	1.459273e+00
surface	1.411605e+00
TerrainParks	1.272134e+00
quad	1.238096e+00

```
Explained Variance Score: 0.7856658038894153
Mean Absolute Error: 5.312114863233078
```

I would recommend that management use Model 2 for predicting ticket price. Although they cannot influence the summit elevation, there are many features in the dataset that are not controllable such as snowfall and days open last year. Eliminating summit elevation is not producing clear improvements in predictive power or insight into potential interventions.